Attention allocation in multi-alternative choice

()

Here we consider the problem of a decision maker who must choose between one of several goods. The decision maker has limited attentional capabilities and is under some time pressure (perhaps only due to opportunity cost). Given these constraints, how should the agent divide her attention among the competing options? Here we derive a near-optimal solution to this problem and compare to human fixation patterns.

Problem statement

An agent must choose between k goods, each having some true utility u_i. The agent does not have knowledge of these exact values, but she can draw samples from k Normal distributions centered on these true utilities. At some point she chooses one of the items and receives a payout equal to the true utiliy of the chosen item.

We assume sampling has a cost, which may be due to explicit time cost, implicit opportunity cost, internal cognitive cost, or some combination of the three. We additionally assume that *changing* the focus of attention is costly; that is, it is more costly to sample from a distribution that was not sampled on the last time step. Because we are interested in modeling eye tracking data, we refer to the item that was most recently sampled from as the *fixated* item; we refer to changing the focus of attention, i.e. sampling from a different distribution as a *saccade.*

POMDP model

We can model this problem as a partially observable Markov decision process (POMDP) which defines a set of states, actions, and observations as well as functions that determine how action and state determine observations and future states. In our model, state defines the true (unknown) utility and sampling precision of each item as well as the (known) attentional state of the agent i.e. which item is fixated. Actions can be attentional (saccading to a new item) or physical (choosing an item). Observations are noisy estimates of the fixated item's utility. Rewards capture the cost of time/attention, the additional cost of saccades, and the utility of the item that is ultimately chosen. Formally, we define a POMDP $(\mathcal{S}, \mathcal{A}, T, R, \Omega, O)$ where

- $\mathcal{S} = \mathbb{R}^k \times \{1, \ldots, k\}$ is the state space. The state is broken up into the true utility of each item $\mathbf{u}$, and the current fixation f.
- $\mathcal{A} = \{\texttt{NOOP}, f_1, \ldots, f_k, c_1, \ldots, c_k\}$ is the action space, where `NOOP` has no effect, f_i saccades to item i, and c_i chooses item i, ending the episode.
- $T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the deterministic transition function that updates only the fixation portion of state based on saccade actions.
- $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function that gives a constant negative reward at each time step, plus an additional negative reward for making saccades. For the choice actions, it gives a reward equal to the utility of the chosen item.
- $\Omega = \mathbb{R}$ is the set of possible observations, which are samples of an item's utility.
- $O : \mathcal{S} \times \mathcal{A} \times \Omega \rightarrow [0, 1]$ is the observation function that gives the probability of drawing a utility sample given the true utility of the fixated item. $O(s, o) = \text{Normal}(o; u_f, \sigma)$ where f is the fixated item and σ is a free parameter that determines how noisy the samples are.

Optimal solution

Following Kaelbling et al. (1998), we break down the problem into two parts: a state estimator and a policy. The state estimator maintains a belief (i.e. a distribution over states) based on the sequence of actions and observations. The policy selects the action to take at each time step given the current belief. By combining these two parts, we create an agent that optimally selects fixations and choices given the full history of previous observations.

State estimator

For ease of exposition, we focus on the belief over item values, noting that the currently fixated item is simply the target of the most recent fixation action. The belief over item values $\mathbf{u}$ is a multivariate Gaussian with mean vector μ and diagonal precision matrix $\Sigma = \text{diag}(\lambda)$. Because the observation o is only informative about the currently fixated item, only the belief about the fixated item changes at each time step. We derive this update by Bayesian inference [cite Murphy?], resulting in

$$
\begin{aligned}
\lambda_f(t+1) &= \lambda_f(t) + \sigma^{-2} \\
\mu_f(t+1) &= \frac{\sigma^{-2} o + \lambda_f(t)\mu_f(t)}{\lambda_f(t+1)} \\
\lambda_i(t+1) &= \lambda_i(t) \text{ for } i \neq f \\
\mu_i(t+1) &= \mu_i(t) \text{ for } i \neq f
\end{aligned}
\tag{1}
$$

The belief is initialized to the prior distribution over (u). For now, we assume that the agent knows the true distribution from which utilities are sampled from. For now, we assume utilities are standard-normal distributed; thus, we have $\mathbf{mu}(0) = 0$ and $\mathbf{lambda}(0) = 1$

Policy

The policy makes two kinds of actions: saccades and choices. We can immediately simplify the problem by reducing the set of choices to a single action, $\perp$, which selects the item that has maximal expected value given the current belief, $\arg\max_i \mu_i(t)$. With this reduction, the problem becomes a *metareasoning* problem (Hay et al. 2012): at each time step the policy must decide whether or not to gather more information and which (if any) item to gather information about. Such problems are generally impossible to solve exactly due to the infinite (continuous) space of possible beliefs. To address this difficulty, Callaway et al. (2018) proposed a reinforcement learning method for identifying metareasoning policies. They found that their method, Bayesian Metalevel Policy Search (BMPS), found near-optimal policies for a bandit-like metareasoning problem with similar structure to the present problem. Thus, we apply their method and treat the identified policy as "optimal".